

# A systematic evaluation of large language models of code

Year: 2022 | Citations: 773 | Authors: *Frank F. Xu, Uri Alon, Graham Neubig*

---

## Abstract

Large language models (LMs) of code have recently shown tremendous promise in completing code and synthesizing code from natural language descriptions. However, the current state-of-the-art code LMs (e.g., Codex) are not publicly available, leaving many questions about their model and data design decisions. We aim to fill in some of these blanks through a systematic evaluation of the largest existing models: Codex, GPT-J, GPT-Neo, GPT-NeoX-20B, and CodeParrot, across various programming languages. Although Codex itself is not open-source, we find that existing open-source models do achieve close results in some programming languages, although targeted mainly for natural language modeling. We further identify an important missing piece in the form of a large open-source model trained exclusively on a multi-lingual corpus of code. We release a new model, PolyCoder, with 2.7B parameters based on the GPT-2 architecture, that was trained on 249GB of code across 12 programming languages on a single machine. In the C programming language, PolyCoder outperforms all models including Codex. Our trained models are open-source and publicly available at <https://github.com/VHellendoorn/Code-LMs>, which enables future research and application in this area. We have an online appendix at <https://arxiv.org/abs/2202.13169>.